Polyploidy underlies co-option and diversification of biosynthetic triterpene pathways in the apple tribe

Wenbing Su[a,1,2], Yi Jing[b,1], Shoukai Lin[c,1], Zhen Yue[b,1], Xianghui Yang[a,1], Jiabao Xu[b], Jincheng Wu[c], Zhike Zhang[a], Rui Xia[a], Jiaojiao Zhu[d], Ning An[d], Haixin Chen[b], Yanping Hong[a], Yuan Yuan[a], Ting Long[a], Ling Zhang[a], Yuanyuan Jiang[a], Zongli Liu[a], Hailan Zhang[a], Yongshun Gao[a], Yuexue Liu[a], Hailan Lin[c], Huicong Wang[a], Levi Yant[e], Shunquan Lin[a,3], and Zhenhua Liu[d,3]

[a]State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops, Ministry of Agriculture and Rural Affairs, College of Horticulture, South China Agricultural University, Guangzhou 510642, China; [b]Research Cooperation Department, Beijing Genomics Institute Genomics, Shenzhen 518083, China; [c]Key laboratory of Loquat Germplasm Innovation and Utilization (Fujian Province), Putian University, Putian 351100, China; [d]Joint Center for Single Cell Biology, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China; and [e]Future Food Beacon and School of Life Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom

Edited by Richard A. Dixon, University of North Texas, Denton, TX, and approved April 5, 2021 (received for review January 28, 2021)

Whole-genome duplication (WGD) plays important roles in plant evolution and function, yet little is known about how WGD underlies metabolic diversification of natural products that bear significant medicinal properties, especially in nonmodel trees. Here, we reveal how WGD laid the foundation for co-option and differentiation of medicinally important ursane triterpene pathway duplicates, generating distinct chemotypes between species and between developmental stages in the apple tribe. After generating chromosome-level assemblies of a widely cultivated loquat variety and *Gillenia trifoliata*, we define differentially evolved, duplicated gene pathways and date the WGD in the apple tribe at 13.5 to 27.1 Mya, much more recent than previously thought. We then functionally characterize contrasting metabolic pathways responsible for major triterpene biosynthesis in *G. trifoliata* and loquat, which pre- and postdate the Maleae WGD, respectively. Our work mechanistically details the metabolic diversity that arose post-WGD and provides insights into the genomic basis of medicinal properties of loquat, which has been used in both traditional and modern medicines.

polyploidy | comparative genomics | evolution | triterpene biosynthesis | apple tribe

Plants generate a vast array of specialized metabolites, which differ by species and tissue. This astounding diversity has long been thought to arise largely by gene duplications, followed by differentiation of gene expression and function between duplicates (1–3). Whole-genome duplication (WGD) contributes to the generation of single gene duplicates and has played important roles in plant genome function and evolution (4–6). Examples for WGD-associated metabolic diversity are known for the innovation of glucosinolates in the Brassicales and for oil biosynthesis in wild olive trees (7–9). In addition to underlying change in enzyme functions, WGD also plays a role in the evolution of metabolic gene expression, which further impacts the biosynthesis of specialized metabolites (3, 10, 11). However, tracing the evolution of a particular metabolic pathway and characterizing functional impacts of WGDs are still challenging, mainly due to the ancient status of relevant WGDs and the formidable complexity of metabolic pathways.

Within the apple tribe there exist medicinally important but understudied species harboring metabolic innovations that we hypothesized were based on the foundation laid by the tribe-specific WGD. A native species from China, loquat (*Eriobotrya japonica* Lindl), has been cultivated as a fruit tree worldwide and used in both traditional and modern medicines (12, 13). Uses include treatment of coughing, documented as early as 1590 CE in the Chinese Encyclopedia of Botany and Medicines [Li Shi-Zhen, Ben Cao Gang Mu (14)] and recent commercial herbal syrups treating throat pain. Recent work showed that ursane-type triterpenes in loquat, that is, ursolic acid (UA) and corosolic acid (CA), are a major class of bioactive compounds with anti-inflammatory, antidiabetic, and anticancer activities (15–17). However, gene pathways encoding the biosynthesis of these bioactive triterpenes have not been identified and characterized in loquat. It has been proposed that members in the apple tribe such as loquat, apple, and pear have been derived from an ancient Rosaceae species via WGD (autopolyploidization), although others have suggested allopolyploidizations between sister species (18, 19). High-quality genome assemblies for species within (apple and pear) (20–22) and outside (strawberry and peach) (23, 24) the apple tribe have been generated, yet genome sequences from the tribe's closest outgroup Gillenieae are still not available.

Here, we merge comparative genomics, transcriptomics, metabolomics, and functional assays to understand how WGD underpinned the diversification of biosynthetic pathways encoding major

Significance

Plants are a primary source of both traditional and modern drugs due to their astounding capability to synthesize diverse molecules. The fruit tree loquat in the apple tribe has been long used in medicine to treat cough, chronic bronchitis, and asthma, yet why loquat—but not its relatives—evolved these medicinal properties is unknown. Here, we generate high-quality genomes of loquat and a relative, which are separated by a recent whole-genome duplication (WGD) around 13.5 to 27.1 Mya. We revealed the post-WGD diversification of triterpene biosynthesis and the exceptionally high levels of bioactive ursane-type triterpenes specifically in loquat. Our work underscores the importance of WGD-associated metabolic diversification underlying the bioactivity of some medicinal plants.

Author contributions: W.S., Y. Jing, Shoukai Lin, Z.Y., X.Y., Shunquan Lin, and Zhenhua Liu designed research; W.S., Y. Jing, Shoukai Lin, Z.Y., and X.Y. performed research; Y.Y., T.L., L.Z., Y. Jiang, Zongli Liu, H.Z., Y.G., Y.L., H.L., H.W., and Zhenhua Liu contributed new reagents/analytic tools; W.S., Y. Jing, Shoukai Lin, Z.Y., X.Y., J.X., J.W., Z.Z., R.X., J.Z., N.A., H.C., Y.H., Shunquan Lin, and Zhenhua Liu analyzed data; and W.S., Y. Jing, L.Y., Shunquan Lin, and Zhenhua Liu wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]W.S., Y. Jing, Shoukai Lin, Z.Y., and X.Y. contributed equally to this work.

[2]Present address: Fruit Research Institute, Fujian Academy of Agricultural Science, Fuzhou 350013, China.

[3]To whom correspondence may be addressed. Email: zhenhua.liu@sjtu.edu.cn or loquat@scau.edu.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2101767118/-/DCSupplemental.

Published May 13, 2021.

triterpenes within and outside of the apple tribe. We first generate chromosome-level assemblies of a widely cultivated loquat cultivar *E. japonica* cultivar (cv.) Jiefangzhong (which produces high levels of ursane-type triterpene) and *Gillenia trifoliata*. This allowed us to first identify a single WGD marking the origin of the apple tribe at 13.5 to 27.1 Mya, much more recent than previously thought (18, 22). We then identified a well-conserved group of enzymatic genes, which we functionally confirm to be responsible for the biosynthesis of bioactive ursane-type triterpenes differentially in species within (apple, pear, and loquat) and outside (*G. trifoliata* and peach) the apple tribe. Our analysis shows how WGD underlaid the diversification of metabolic pathways responsible for major triterpene biosynthesis. In addition, WGD provided a basis for the evolution of functionally critical gene expression differentiation and co-option for biosynthesis of triterpenes. Taken together, our results provide a clear example of the post-WGD diversification of an important metabolic pathway and identify the genomic basis for exceptionally high levels of bioactive triterpenes in loquat, which has long been used in both traditional and modern medicines.

Results

Chromosome-Level Assembly of Loquat and *G. trifoliata* Genomes, Which Are Separated by a WGD. To achieve a chromosome-level assembly for loquat (cv. "Jiefangzhong"), we used a combination of single molecule real-time (SMRT) sequencing by PacBio, Hi-C, and Illumina short-read sequencing (*SI Appendix*, Table S1). We first generated 113× coverage (~91.4 Gb) of PacBio long reads for a primary assembly. We then polished with 102× coverage (82.6 Gb) of Illumina short reads using Pilon (25). This resulted in an assembly with a contig N50 of 3.98 Mb and a total length of 760.98 Mb. Next, we generated 97× coverage (78.27 Gb) of Hi-C data, which allowed >96.93% of the contigs (737.64 Mb) to be anchored onto 17 pseudomolecules with a scaffold N50 of 43.16 Mb (*SI Appendix*, Fig. S1*A*). The final genome assembly is ~761 Mb, corresponding to 94.7% of genome size (803 Mb) estimated by flow cytometry (*SI Appendix*, Tables S2 and S3). Our results confirmed that loquat, along with other Maleae species, has a basic chromosome number of $\chi = 17$ (22, 26). However, most of the other Rosaceae have approximately half the base chromosome number of Maleae, for example, species in Rosoideae ($\chi = 7$) (27), Amygdaleae ($\chi = 8$) (24, 28), and Gillenia ($\chi = 9$) (29). It has been suggested that ancestors of Maleae were derived from allopolyploidization between species with basic chromosome numbers of eight and nine, respectively (19). However, phylogenomic analysis suggests that the extant Maleae species resulted from a within-species WGD (autopolyploidy), based on intra- and intergenomic synteny analysis across Maleae (i.e., apple and pear) and outgroup species (22, 26). So far, comparative genomic analyses have not included species with a basic chromosome number of $\chi = 9$. We thus sequenced and assembled the genome of a Gillenieae species, *G. trifoliata,* using the same methods described above, generating 898× coverage (~287.5 Gb) of PacBio long reads, 130× coverage (~41.4 Gb) of Illumina short reads, and 130× coverage (~37.4 Gb) of Hi-C data (*SI Appendix*, Table S1). These resulted in an assembly ~280.76 Mb with a contig N50 of 828 kb, corresponding to 87.7% of genome size (320 Mb) estimated by flow cytometry (*SI Appendix*, Tables S2 and S3). The majority (96%) of the contigs were anchored into nine pseudochromosomes with a scaffold N50 of 30.09 Mb according to Hi-C analysis (*SI Appendix*, Fig. S1*B*). To assess the quality of the assembled genomes, we performed Benchmarking Universal Single-Copy Orthologs (BUSCO) and RNA transcripts mapping analysis (*Materials and Methods*). Over 97% of BUSCO complete genes can be detected in both genomes. In addition, over 98% of RNA transcripts can be mapped onto both genomes (*SI Appendix*, Table S3). The high quality of the two genome assemblies thus enabled high-confidence annotation of both coding and noncoding genomic regions (*SI Appendix*, Table S4).

To determine species groupings in relation to the Maleae WGD, we used 661 single-copy orthologs from 10 related species for phylogenetic analysis. This revealed loquat grouping with pear and apple and separated from *G. trifoliata*, peach (*Prunus persica*), and other Rosaceae species (Fig. 1*A*). Molecular clock analysis indicated that the apple tribe (Maleae) diverged from Gillenieae between 13.5 and 27.1 Mya (Fig. 1*A*), suggesting that the emergence of the apple tribe was much more recent than previously thought (18, 22). Using intra- and intergenomic syntenic analysis, we observed a 2:1 syntenic depth ratio when comparing loquat with the species (*G. trifoliata* and *P. Persica*) that diverged before the split with Maleae but showed a 1:1 syntenic depth ratio within Maleae (apple versus loquat) (Fig. 1*B*). Furthermore, the spectrum of synonymous substitutions per synonymous site (Ks) of these syntenic blocks confirmed a single WGD peak (Ks around 0.16) for the two Maleae species (loquat and apple) but not for *G. trifoliata* and the outgroup species grape (Fig. 1*C*). In line with this, ancestral reconstruction analysis indicated that the genomes of extant apple tribe were duplicated from a common ancestor with nine chromosomes, despite extensive genome reorganizations after the WGD (*SI Appendix*, Fig. S2). Altogether, our genome assembly of loquat and *G. trifoliata* provided evidence for a WGD in a common ancestor of the apple tribe, after separating from the Gillenieae 13.5 to 27.1 Mya (Fig. 1*A*).

Loquat Leaves Accumulate Exceptionally High Levels of Ursane-type Triterpenes. As triterpenes have been suggested as major bioactive compounds in loquat (13, 30), we next set out to investigate the triterpene profiles in species evolved before and after the emergence of the apple tribe and the associated WGD. Young and old leaves from loquat and apple (within Maleae) and peach and *G. trifoliata* (outgroup) were analyzed by gas chromatography mass spectrometry (GC-MS). Pentacyclic triterpenes including α-amyrin (AA), β-amyrin (BA), UA, oleanolic acid (OA), maslinic acid (MA), and CA were identified as major triterpenes in leaves (Fig. 2*A*). Strikingly, loquat's old leaves accumulated ~10-fold higher triterpenes levels than any other species (Fig. 2*B*). Of these, ursane-type triterpenes (derived from AA) were found dominantly accumulated in loquat (about three times more highly accumulated than oleanane-type triterpenes) (Fig. 2*B* and *SI Appendix*, Fig. S3*A*). To gain a higher-resolution triterpene profile in loquat, we selected leaves from five developmental stages for GC-MS. Our results showed that older leaves accumulated much higher levels of a broader triterpene array. UA was constantly detected since stage 2, whereas CA was largely appeared only in stage 5 (mature old leaves) (Fig. 2 *C* and *D* and *SI Appendix*, Fig. S3 *B–D*). Notably, leaves from stage 5 have been commonly used in folk medicines (31).

Characterization of Genes Encoding the Biosynthesis of Major Triterpenes in Loquat. The first committed step for biosynthesis of plant triterpenes is catalyzed by oxidosqualene cyclases (OSCs), which fold the linear substrate 2,3-oxidosqualene into >200 diversified structures (32). To characterize candidate genes encoding the biosynthesis of major triterpenes in loquat, we systematically mined OSC genes in the available genomes from Maleae (loquat, apple, and pear) and two outgroup species (peach and *G. trifoliata*) (*Materials and Methods*). These analyses identified 8, 6, 14, 16, and 13 OSC genes in peach, *G. trifoliata*, apple, pear, and loquat, respectively (Fig. 3*A*). To better understand triterpene biosynthetic profiles before and after diversification of the Maleae, we examined general gene expression profiles for *OSCs* in peach (33), *G. trifoliata* (this study), apple (26), pear, and loquat (this study). This analysis first revealed that gene expression for OSC genes were unevenly distributed across phylogenetic clades, with a subclade of triterpene-related OSC genes showing dramatically higher gene expression levels relative to other OSC genes (hereafter, referred to as main subclade) (Fig. 3*A*). In this main subclade, two

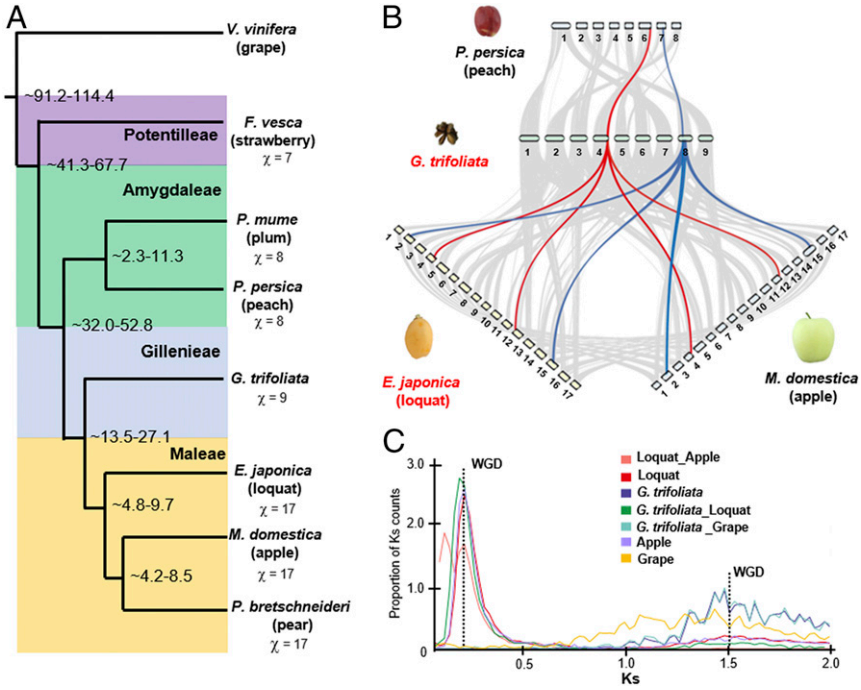

Fig. 1. Comparative analysis including newly assembled genomes of loquat and *G. trifoliata* identified a single WGD at the base of Maleae ~13.5 to 27.1 Mya. (*A*) Phylogenetic relationships of loquat and other Rosaceae species. The ML phylogeny was inferred by 661 single-copy orthologs. Species split time based on molecular clock analysis is indicated on each branching node. (*B*) Intergenomic syntenic analysis between peach, *G. trifoliata*, apple, and loquat. Genomic regions in peach and *G. trifoliata* could be aligned with up to two regions in loquat and apple. Two examples in chromosomes 4 and 8 in *G. trifoliata* are highlighted by colors. (*C*) Ks distributions for gene pairs from syntenic blocks of *G. trifoliata*, grape, apple, and loquat. The two predicted WGD events are indicated by dashed lines. Ks around 1.5 indicates the gamma WGD occurred at the base of the eudicots. Ks around 0.16 indicates the more recent WGD specifically occurred in the apple tribe.

OSCs from apple have been previously characterized as AA or BA synthase, which generate backbones for ursane-type and oleanane-type triterpenes, respectively (34). Therefore, we hypothesized that the two loquat OSC genes (*Ej00015273* named as *EjOSC1* and *Ej00096061* named as *EjOSC2*) identified from the main subclade were likely responsible for the biosynthesis of major triterpenes in loquat.

As major triterpenes detected in loquat leaves are chemical derivatives of AA and BA, in order to identify downstream genes, we next used a weighted gene coexpression network analysis (WGCNA) to identify genes that were highly correlated to *EjOSC1* and *EjOSC2*. A total of 29 modules were detected (Dataset S1). The two highly expressed *OSCs* were both located in the same module. To identify potential triterpene-modifying genes, we ranked the genes according to correlation weights to *EjOSC1* and *EjOSC2*, respectively. This analysis identified three and four tightly associated enzymatic genes for *EjOSC1* and *EjOSC2*, respectively (Fig. 3*B* and Dataset S2). Phylogenetic analysis of those genes revealed that four CYP716 genes, which were previously identified as pentacyclic triterpene-modifying P450 genes (34, 35), were included in the correlation network (*SI Appendix*, Fig. S4 and Dataset S2).

To test the enzymatic functions of those associated *OSC* and *CYP716* genes, we cloned all six unique genes from both top correlated–enzymatic gene lists above and expressed them transiently in *Nicotiana benthamiana* leaves. Expression of *EjOSC1* resulted in a mixture of AA:BA (95:5) (Fig. 3*C* and *SI Appendix*, Fig. S5*A*). Further coexpression of *EjCYP716A1* resulted in the formation of a mixture of UA:OA (85:15) (Fig. 3*C* and *SI Appendix*, Fig. S5*B*), suggesting that EjCYP716A1 was able to catalyze a three-step oxidation reaction at C-28 on AA and BA to produce UA and OA. Further expression of *EjCYP716C1* gave rise to a mixture of CA:MA (81:19) (Fig. 3*C* and *SI Appendix*, Fig. S5*C*), suggesting that EjCYP716C1 was able to catalyze the C-2α hydroxylation of OA and UA to produce MA and CA. The higher level of ursane-type triterpenes (UA and CA) in these combinational assays was consistent with metabolite profiles in loquat leaves examined (*SI Appendix*, Fig. S2 *A* and *C*). Moreover, MA and CA (produced by CYP716C) were accumulated in later stages compared to UA and OA (produced by CYP716A) (Fig. 2*C*), suggesting that in vivo biosynthesis of major triterpenes in loquat is likely through the flux of OSC-CYP716A-CYP716C (Fig. 3*D*). In parallel, we tested the second set of correlated genes (*EjOSC2-EjCYP716A2-EjCYP716C2*) in the *N. benthamiana* leaf system. A similar pattern for metabolite production was observed (Fig. 3*C* and *SI Appendix*, Fig. S5), suggesting that both of these two sets of enzymes are responsible for the biosynthesis of major triterpenes in loquat. Genes encoding for MA and CA production in loquat were also characterized in distant orders of plant species (35–37), suggesting that the OSC-CYP716A-CYP716C pathway for triterpene biosynthesis is widespread in plants.

Evolution of Major Triterpene Biosynthesis before and after Emergence of the Apple Tribe. To understand the evolution of these two key functional modules, we first looked at OSCs, which catalyzed the first committed step in the pathway. In the main subclade, a single copy of *OSC* presented in the species evolved before the WGD whereas two copies were found in all the Maleae species examined (Fig. 3*A*). We cloned the single-copy *Gt00028126* from *G. trifoliata* in the main subclade. Transient expression of *Gt00028126* in *N. benthamiana* gave a mixture of AA and BA at a ratio of 92:8 (*SI Appendix*, Fig. S6 *A* and *B*), similar to the functionally characterized *OSCs* in loquat (this study) and apple (34), suggesting the *OSCs* from the main subclade shared a common ancestor. Syntenic analysis further revealed that the duplicated *OSCs* in the apple tribe were derived from the WGD event (Fig. 4*A*). Interestingly,

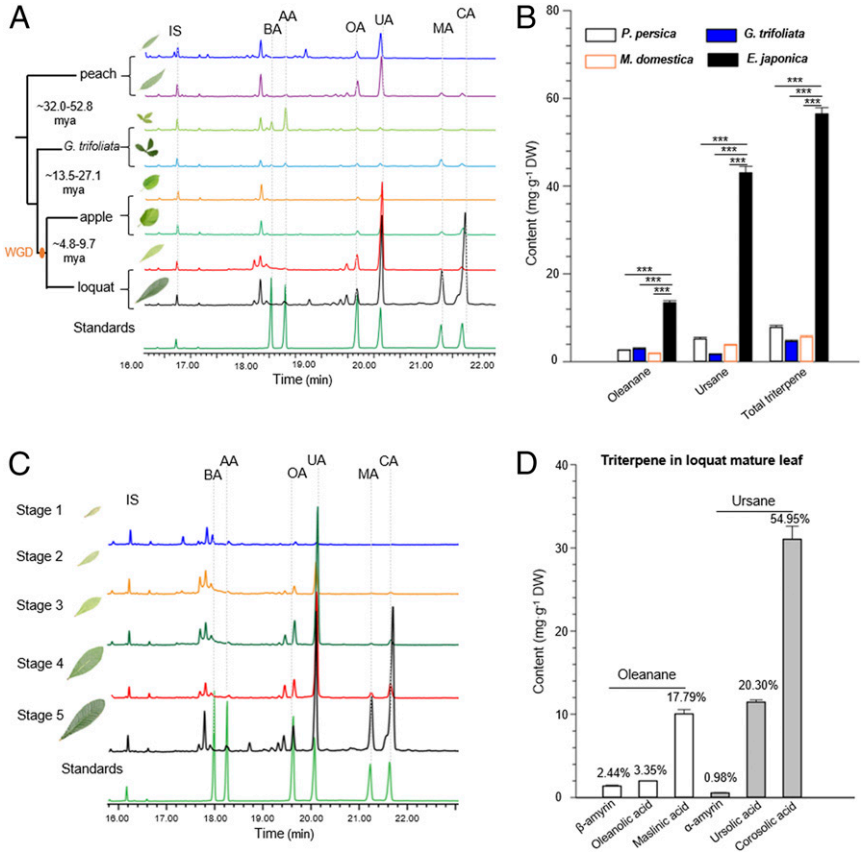

Fig. 2. Loquat leaves accumulate exceptionally high levels of ursane-type triterpenes. (*A*) GC chromatograms show major pentacyclic triterpenes in young and mature leaves of loquat, apple, *G. trifoliata*, and peach. IS, internal standard (Coprostanol); BA, beta-amyrin; AA, alpha-amyrin; OA, oleanolic acid; UA, ursolic acid; MA, maslinic acid; and CA, corosolic acid. (*B*) Triterpene content in mature leaves of loquat, apple, *G. trifoliata*, and peach. Error bars present means ± SE, with three biologically independent replicates; *** presents significant difference at $P < 0.001$ by two-sided Student's *t* test. (*C*) GC chromatograms show major triterpene profiles in loquat leaf developmental stages. (*D*) Major triterpene content in mature loquat leaves ($n = 3$). Error bars present means ± SE with three biologically independent samples.

our syntenic analysis also identified a tandem *OSC* duplicate, which was likely coevolved with the *OSCs* in the main subclade (Fig. 4*A*). Phylogenetic analysis (Fig. 3*A*) also supported that they shared a common ancestor. However, the tandem duplicate-derived *OSCs* all showed a low level of gene expression (Fig. 3*A*) and are retained as a single copy after the WGD in the apple tribe (Fig. 4*A*).

Using similar approaches, we found that the functionally characterized *CYP716Cs* (encoding C2-hydroxilase) were also derived from a pair of ancestral tandem duplicated genes and retained as four copies in apple and loquat after the WGD (Fig. 4*B*). However, this seems to not be the case for the functionally characterized *CYP716As* (encoding C28-oxidase). Synteny analysis showed that *EjCYP716A1* was likely evolved as early as in an ancestor predating the emergence of peach. In line with this, functional characterization of a syntenic ortholog (*Gt00010673*) in *G. trifoliata* revealed the conserved enzymatic function of this gene to *EjCYP716A1* (*SI Appendix*, Fig. S6). However, syntenic evidence supporting ancestral tandem duplication of CYP716A genes was not found (Fig. 4*B*). In agreement with this, *EjCYP716A2* was found located in a different chromosome and thus most likely experienced relocation after the WGD (Fig. 4*D*). Interestingly, although *Ej00014855*, the *EjCYP716A1* WGD-associated duplicate copy (Fig. 4*C*), did not show a concerted coexpression pattern with *EjOSC1* and *EjCYP716C1* (Fig. 3*B*), it showed similar enzymatic functions as EjCYP716A1 when tested in *N. benthamiana* (*SI Appendix*, Fig. S6*C*). We therefore interpreted that this gene might contribute to triterpene variations between developmental stages in the leaf tissue (Fig. 2*C*).

To reconstruct the evolutionary status for biosynthesis of major triterpenes before the WGD, we next applied WGCNA using transcriptome data generated from nine developing tissues in *G. trifoliata*. *G. trifoliata OSC-CYP716A-CYP716C* genes, for which their orthologs have been functionally characterized in loquat, appeared in a highly correlated expression network (Fig. 5 and Dataset S3). However, their degrees of correlation (reflected by ranking distance to OSC gene) were much lower in comparison to relationships identified for genes in loquat using the same approach (Fig. 5 and Dataset S4). This indicates that the major triterpene biosynthetic pathway genes gained concerted gene expression following the WGD, contributing to especially high levels of detected triterpenes in loquat (Fig. 5).

Discussion

Plants are remarkable chemists, estimated to synthesize approximately one million specialized metabolites (38), yet these differ dramatically in terms of composition and quantity between species, as well as across development. The genomic basis of this astounding metabolic complexity has been long proposed to arise via gene duplication, followed by subfunctionalization and neofunctionalization (39–41). Recently, genomic reorganizations, such as formation of diterpene and triterpene gene clusters, are also found as new genomic features associated with metabolic diversification in plants (42–45). There are a wealth of cases

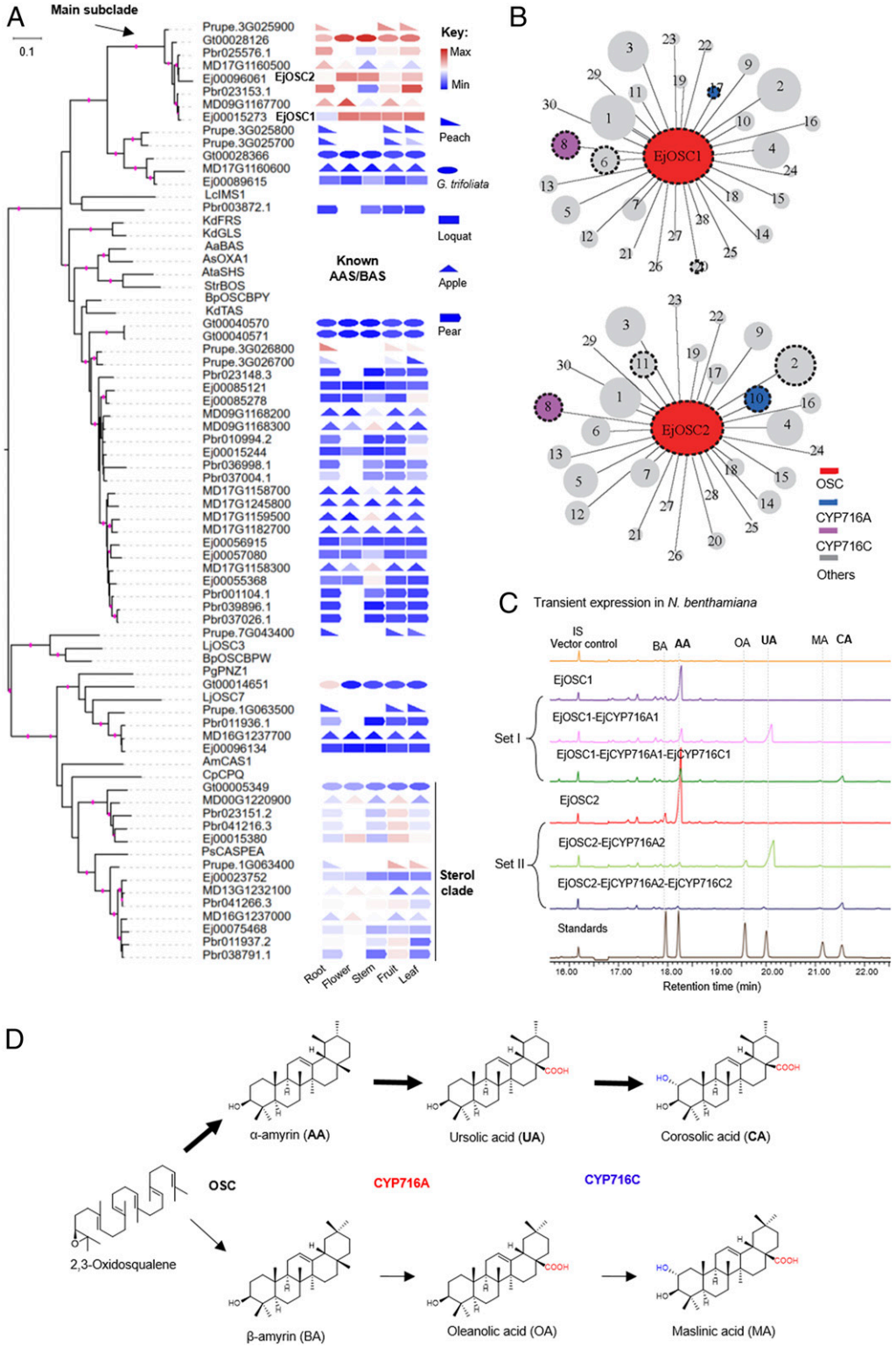

Fig. 3. Characterization of candidate genes encoding biosynthesis of major triterpenes. (*A*) Phylogenetic relationship and gene expression profile of OSCs from available Maleae and outgroup species. The ML tree was inferred with OSC proteins. Tree node support (>80%) is indicated by pink dots on branches (1,000 bootstrap replicates). The main subclade with higher gene expression levels was indicated by an arrow. Notably, gene expression levels are comparable within species only. Gene expression levels based on transcript per million values are indicated as a heatmap. (*B*) *EjOSC* centric coexpression networks. The top 30 correlated genes were shown in the network. The genes in dashed circles were predicted to encode enzymes. Detailed gene list can be found in Dataset S2. (*C*) GC-MS chromatograms of major triterpenes extracted from *N. benthamiana* leaves transiently expressed with candidate genes. (*D*) The pathway for biosynthesis of major triterpenes in loquat. The thick arrows present dominant flux toward ursane-type triterpenes due to promiscuous functions of OSC, CYP716A, and CYP716C enzymes.

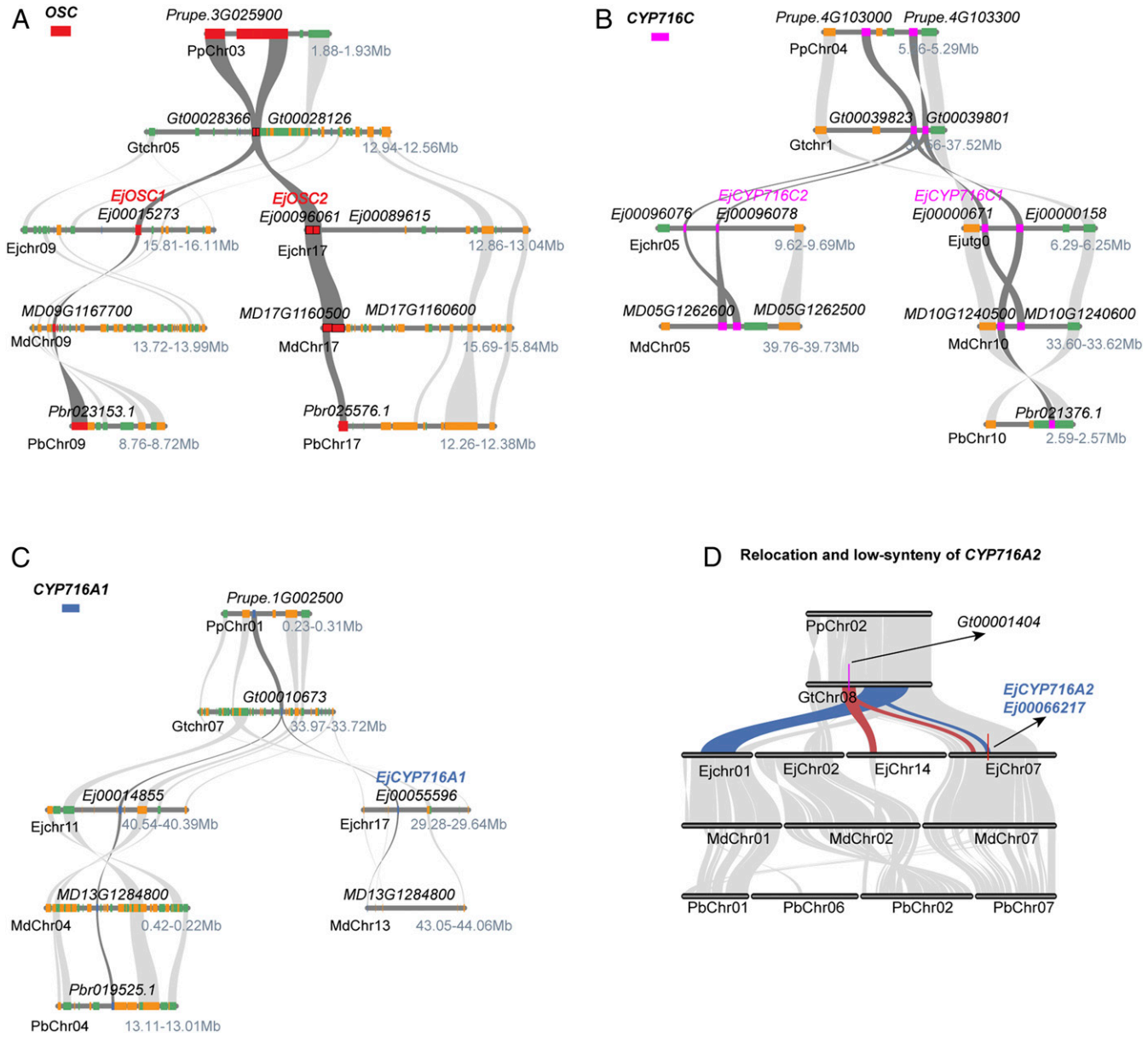

Fig. 4. Syntenic analysis of genes encoding for biosynthesis of major triterpenes in the apple tribe. (*A–C*) Microsynteny across peach, *G. trifoliata*, loquat, apple, and pear for *EjOSC1* and *EjOSC2* in *A*, for *EjCYP716C1* and *EjCYP716C2* in *B*, and for *EjCYP716A1* in *C*. Syntenic blocks are connected by lines. (*D*) Macrosynteny across peach, *G. trifoliata*, loquat, apple, and pear for *EjCYP716A1*. *Gt00001404* and *EjCYP716A2* formed a monophyletic branch in Fig. 3*A*, indicated by arrows.

tracing single gene duplication events to address the contribution of these local variants on diversification of metabolites (41, 46, 47). However, the impact of WGD-associated metabolic diversification is largely unknown, despite the obvious prevalence of WGD throughout plant evolution.

To address this, we carried out functional and genomic analyses of the evolution of a triterpene pathway before and after a single WGD event at the base of the apple tribe. Our biochemical analysis showed that major triterpenes in loquat and *G. trifoliata* were synthesized by three promiscuous enzymes, OSC, CYP716A, and CYP716C, with a dominant function catalyzing ursane-type triterpenes (Fig. 2*D*). This suggests that the triterpene profile variation between species is mainly defined by the dynamic catalytic activity of those enzymes. It also indicates that triterpene abundance variation between *G. trifoliata* and loquat is likely due to the expansion of gene duplicates following WGD. Our systematic analysis of genomes, transcriptomes, and metabolites across species before and after the apple tribe WGD allowed reconstruction of the evolutionary history of major triterpenes biosynthesis and a view into the evolutionary mechanism driving the differences between these species. Before the WGD, genes for biosynthesis of major triterpenes were indeed present and are likely functionally co-opted via tandem (for *OSC* and *CYP716C*) or single gene duplications (for *CYP716A*). This nascent functional module is present in *G. trifoliata*, where low levels of triterpenes were detected. Following WGD, the pathway was fully duplicated and additional genes derived from ancestral tandem or general gene duplications were retained in loquat. Consequently, genes for biosynthesis of major triterpenes were markedly expanded. In addition, OSC genes in loquat (*EjOSC1* and *EjOSC2*) showed distinct association with *CYP716A* and *CYP716C* genes in WGCNA analysis (Fig. 5), suggesting that they are undergoing a process of differential gene

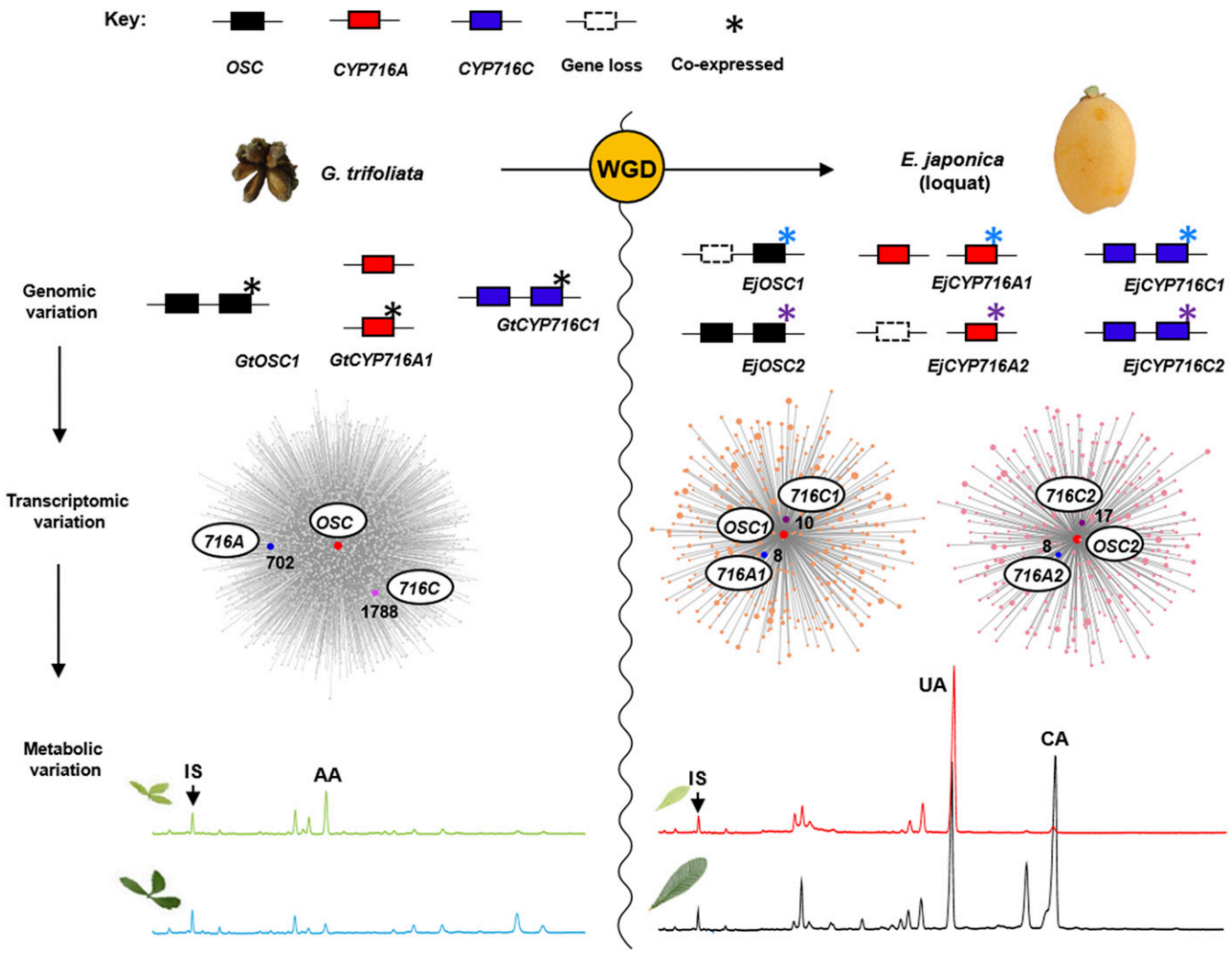

Fig. 5. WGD-associated diversification of the metabolic pathway for biosynthesis of major triterpenes before and after the emergence of the apple tribe. Pathway genes in the OSC-centric coexpression network are indicated by colors. Numbers represent correlation rankings based on WGCNA weights to targeted OSCs (reference Datasets S2 and S4). IS, internal standard; AA, alpha-amyrin; UA, ursolic acid; and CA, corosolic acid.

expression pattern between the duplicated pathways. This interpretation is further supported by the fact that the first set of genes (*EjOSC1-EjCYP716A1-EjCYP716C1*) were highly expressed in old leaves, whereas the second set of genes (*EjOSC2-EjCYP716A2-EjCYP716C2*) showed the opposite trend (*SI Appendix*, Fig. S7).

Taken together, our results suggest that WGD laid the foundation for expansion of metabolic pathways, then allowing natural selection to drive the co-option and differentiation of triterpene producing machineries in loquat. Our genomics-driven approach thus provides an explicit example of WGD-associated metabolic diversification. It also sheds insight into the genomic basis underpinning medicinal properties of loquat that has been long used in traditional medicines.

Materials and Methods

Plant Material and Growth Conditions. *E. japonica* cv. Jiefangzhong was used for genomic and transcriptomic analysis. Loquat and peach trees were grown in the Loquat Germplasm Resources Garden in South China Agricultural University. Apple trees were grown in the Botanic Garden in Guangzhou Academy of Agricultural Sciences (Guangzhou, China). Seeds of *G. trifoliata* were provided by the Missouri Botanical Garden (United States). *G. trifoliata* seeds were sterilized in 10% sodium hypochlorite for 8 min and rinsed with sterile water three times. The seeds were then immersed in 0.02% GA_3 to promote germination for 1 d. Germinated seeds were cultivated in propagation medium (MS with 1.5 mg · L^{-1}TDZ and 0.4 mg · L^{-1} IBA) at 25°C for 28 d. *N. benthamiana* were planted under long-day conditions (16 h light/8 h dark) at 22°C.

Genome Sequencing. Loquat young leaves (developmental stage 2) were used for genomic DNA extraction as described (48). PacBio library (20 kb) was constructed and sequenced using 10 SMRT cells on a PacBio Sequel. A total of ~91 Gb SMRT reads with an average length of ~11 Kb were generated. Eight size-selected Illumina genomic libraries ranging from 350 bp to 40 kb were constructed and sequenced on an Illumina HiSeq2000 system. Hi-C libraries were prepared as previously described (49) and sequenced on BGISEQ-500 system with PE 100. A total of ~78.27 Gb reads were generated. For *G. trifoliata,* 10 g young leaves were collected and frozen in liquid nitrogen before DNA extraction. One PacBio 20-kb library was constructed and sequenced using 1 SMRT cells on a PacBio Sequel. A total of ~287.54 SMRT reads with an average length of 23.56 Kb were generated. Two Illumina genomic libraries (500bp and 5kb) were constructed and sequenced on an Illumina HiSeq 2000 system, which generated 37.9 Gb (115×) and 3.5 Gb (10.9×) data, respectively. Hi-C library (350 bp) was prepared as previously described (49) and sequenced on Illumina HiSeq X with PE 150. Statistics of sequencing data for the two genomes are listed in *SI Appendix*, Table S1.

Flow Cytometry Analysis. Chopped young leaf tissues were incubated in ice-cold LB01 lysis buffer (15 mM Tris, 2 mM EDTA, 20 mM NaCl, 80 mM KCl, 0.5 mM spermine tetrahydrochloride, 1% β-mercaptoethanol, 0.5% Triton X-100,

2% PVP-40, and 50 μg · ml^{-1} RNase, pH 7.5) for 30 mins to release nuclei. Tomato leaf tissue with known nuclei DNA content was used as size control. Isolated nuclei were stained with 50 μg/mL propidium iodide (PI) and incubated for 30 min at 4 °C in the dark. PI excitation was achieved at 488 nm and emission was analyzed at 675 nm using Beckman-Coulter FC-500 and Becton Dickinson FAC SMeldoy analyzers to determine DNA content as described (50). *Solanum lycopersicum* cv. Stupicke (gifted by Professor Dolezel, Institute of Experimental Botany, Czech Republic) and human leukocytes were used as internal reference standards for *G. trifoliata* and loquat, respectively. Leukocyte and tomato were used as reference standards for having reasonably larger genome size than loquat and *G. trifoliata* and could be easily separated from the tested samples. The leukocyte samples were deidentified prior to use in this study. PI fluorescence was activated by laser emitting at 532 nm on flow cytometry with four biological replicates. Statistics of flow cytometry analysis for the two genomes are shown in *SI Appendix*, Table S2.

Genome Assembly.

Loquat. We assembled the loquat genome by integrating PacBio long-read sequencing, Illumina paired-end, and Hi-C sequencing data. Briefly, the PacBio raw reads were corrected by Canu (51) with the following parameters: minReadLength >3,000 and minOverlapLength >500. The longest 50× Canu corrected reads were then assembled by SMARTdenovo (52) with k-mer (k) =17. Illumina short reads were preprocessed by removal of sequences from bacteria, adapters, low quality, and duplicate reads using SOAPnuke (53) (version 1.6.5) with the following parameters: -n 0.01 -l 20 -q 0.1 -i -Q 2 -G -M 2 -A 0.5 -d. The primary assembly was polished by Illumina short reads using Pilon with default parameters (25). To process the Hi-C raw data, Juicer pipeline (54) was used to align fastq reads to the genome, giving rise to duplication-free data. The resultant sequences and scaffolds were transferred to three-dimensional DNA (55). Briefly, scaffolds <15 Kb were first discarded. To eliminate misjoins from the input scaffolds, two iterative steps were performed. Each step started with a scaffolding algorithm to order and orient the input scaffolds, followed by misjoin corrections to detect errors in the scaffold pool. The edited scaffold pool was used as an input for the next iteration of the misjoin correction algorithm. After the iterations were completed, a single "megascaffold," which concatenates all the chromosomes, was retained for postprocessing. The postprocessing included four steps: 1) a polishing algorithm, which was required for genomes in the RabI configuration; 2) a chromosome splitting algorithm, which was used to extract the chromosome-length scaffolds from the megascaffold; 3) a sealing algorithm, which was required to detect false positives in the misjoin correction process and restore the erroneously removed sequence from the original scaffold; and 4) a merge algorithm, which was used to correct misassembly errors due to undercollapsed heterozygosity in the input scaffolds.

G. trifoliata. We assembled the *G. trifoliata* genome by integrating PacBio long-read, Illumina paired-end, and Hi-C sequencing data. PacBio raw reads, Illumina paired-end, and Hi-C sequencing data were corrected and assembled into assembly using the same method as for loquat. Statistics of assembly data for the two genomes are listed in *SI Appendix*, Table S3.

Transcriptome Sequencing and Assembly. Total RNAs were isolated using RNeasy Plant Mini Kit (QIAGEN). Briefly, ca. 100 mg tissue powder was treated with 450 μL buffer RNA lysis buffer at 56 °C for 3 mins. The extracts were then transferred to a QIAshredder spin column and centrifuged for 2 mins at 12,000 rpm. Ca. 400 μL supernatant were taken, mixed with 200 μL ethanol, and transferred to an RNeasy spin column for RNA binding. The column membrane was washed with 500 μL buffer RNA wash buffer 1 and 700 μL buffer RNA purification ethanol. After drying the column, 30 μL RNase-free water was used to elute the RNA. Complementary DNAs (cDNAs) were synthesized with the PrimeScript RT reagent Kit (Takara). Total RNAs (700 ng) were treated by 1.0 μL genomic DNA Eraser with 2.0 μL Eraser buffer at 42 °C for 5 min. For inverse transcription, 4 μL 5× PrimeScript buffer, 4 μL RNase-free ddH_2O, 1 μL PrimeScript RT Enzyme MixI, and 1 μL Oligo dT Primer were added into the above RNA mix. The mixture was incubated at 37 °C for 30 mins, then 85 °C for 1 min to synthesize the first strain of cDNA. RNA sequencing (RNA-seq) libraries were constructed from cDNAs (insert size of 200 to 350 bp) according to Illumina's protocols.

The cDNA libraries were sequenced on the Illumina's HiSEq. (4000) system. The raw reads were preprocessed by removal of contamination sequences from bacterial, adapters, low quality, and duplicate reads using SOAPnuke (version 1.6.5) with default parameters. De novo transcriptome assembly was performed using trinity (56) with the following parameters: –group_pairs_distance 280–no_version_check–full_cleanup–verbose–min_contig_length 150–CPU 8–min_kmer_cov 3–min_glue 3–bfly_opts '-V 5–edge-thr = 0.1–stderr.

Genome Annotation for Loquat and *G. Trifoliata*.

Annotation of transposable elements. We combined de novo and homology-based approaches to identify transposable elements (TEs). The de novo prediction was carried out using RepeatModeler (57) to construct a repeat library with default parameters. This library was used as a database for RepeatMasker (57) to find and classify repeats. Homology prediction was carried out using RepeatMasker or RepeatProteinMask by aligning the genome sequence to the Repbase version 21.12 database (58) to identify TEs. All the repeat sequences identified by different methods were combined as the final list for annotation. Tandem repeats were searched using Tandem Repeats Finder (59).

Annotation of noncoding RNAs. The tRNAScan-SE (version 1.3.1) (60) algorithm with default parameters was applied to predict transfer RNA (tRNA) genes according to the structural characteristics of tRNA. BLAST (61) was used to identify ribosomal RNA (rRNA) by aligning with rRNA template sequences (Rfam database version 12) with an E-value of 1e−5. MicroRNA and small nuclear RNA genes were predicted using INFERNAL (http://infernal.janelia.org/) with the Rfam database version 12 (62).

Annotation of genes. Gene structure prediction was conducted by MAKER pipeline (63) integrating with ab initio gene predictions, transcript evidence (de novo assembled transcripts), and protein homology evidence from *Arabidopsis thaliana*, *Malus domestica*, *Pyrus bretschneideri*, *P. persica*, and *Rosa chinensis* after transposons were masked from genome sequences (*SI Appendix*, Table S4). De novo assembled transcripts were used as training data in ab initio prediction software SNAP (64) and AUGUSTUS (65). The MAKER pipeline was processed using trained hidden Markov model profiles. Low-confidence predictions (annotation edit distance <0.4) were filtered, and 43,996 gene models were obtained. Gene function annotations were performed by aligning the proteins of each gene to SwissProt, Gene Ontology (GO), Translated sequence in the European Molecular Biology Laboratory (TrEMBL) database, Clusters of Orthologous Groups (COG), Nonredundant Protein Sequence Database (Nr), and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases with Blast (E-value ≤ 1e-5). The best hit was assigned to each gene. The motifs and domains in protein sequences were annotated using InterProScan version 5.16-55.0 (66) via searching public databases. GO terms for each gene were assigned using Blast2GO (67).

Assessment of Genome Quality. BUSCO (version 3.0.2) (68) with 1,375 single-copy orthologs was used to assess the completeness of the genome assembly. BLAT (69) was used to map the transcripts back to the final assembled genome for loquat and *G. trifoliata*, respectively (*SI Appendix*, Table S3).

Estimation of Gene Expression Level. The cleaned RNA-seq reads (taken from above de novo transcriptome assembly) were aligned to the gene set by Bowtie (70) (version 2.2.5), and gene expression levels were estimated by RNA-seq by Expectation Maximization (RSEM) (71) (version 1.2.12). Expression level heatmap based on transcripts per million (TPM) values was drawn using TBtools (72). For apple and peach, the transcriptome data of five tissues (root, stem, leaf, flower, and fruit) and three tissues (leaf, root, and fruit) were downloaded from National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) (Accession numbers are listed in *SI Appendix*, Table S5.)

Phylogenetic Analysis. Proteins from 10 species (*E. japonica*, *G. trifoliata*, *M. domestica*, *P. bretschneideri*, *P. persica*, *Prunus mume*, *Fragaria vesca*, *Vitis vinifera*, *A. thaliana*, and *Oryza sativa*) were first mined by BLASTP (61) version 2.2.26 with default parameters to generate pairwise protein sequence with an E-value cutoff of 1e−5. OrthoMCL (73) version 1.4 was used to assign orthologous gene clusters with the following parameters: –pv_cutoff 1–mode 3–abc -I 1.5. We found 661 single-copy orthologous gene families in 10 species (Dataset S5). Those sequences were extracted and aligned using muscle (version 3.8.31) (74). Poorly aligned positions and divergent regions of the alignment were eliminated using trimAl (75). Phylogenetic analysis was performed using a maximum likelihood (ML) method implemented in RaxML (version 8.2.9) (76) with the GTRGAMMA substitution model and 1,000 bootstrap replicates. *O. sativa* was used as the outgroup.

Divergence Time Estimation. Species divergence time based on the above mentioned 661 single-copy orthologous genes was estimated using Markov chain Monte Carlo (MCMC) algorithm from PAML 4.9 (77). The divergence

time of Maleae and prunus was set to ~30 to 61 Mya, the divergence time of *A. thaliana* and *O. sativa* was set to ~148 to 173 Mya, and the divergence time of *A. thaliana* and *V. vinifera* was set to ~106.0 to 119.3 Mya (http://www.timetree.org/) (78). Overall substitution rate was assessed using baseml in PAML 4.9 (77) by setting a REV substitution model. For MCMC analysis, 20,000 iterations were simulated with a sampling frequency of 5,000.

Synteny Analysis. The SynMap tool in the online Co-Ge Platform (https://genomevolution.org/CoGe/) was first used to capture the syntenic blocks within and between genomes. A putative syntenic region included at least four colinear genes. MCScan (python version) was also used to complement the analysis (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version).

Ks Analysis. Intra- and intersyntenic gene pairs were used to calculate Ks values in Fig. 1*C*. Protein sequences from the gene pairs were aligned using MUSCLE (74). Alignments were then translated into coding nucleotide sequences using an in-house Perl script. Ks values based on the alignments of coding nucleotide sequences were calculated using yn00 in PAML 4.9 (77).

Ancestral Reconstruction of Maleae Gene Content. Syntenic blocks of Maleae from the most recent WGD and species divergence were obtained according to Ks value (Ks < 0.55 for intrasyntenic blocks in loquat, apple, and pear; Ks < 0.45 for intersyntenic blocks of loquat–apple, loquat–pear, and apple–pear; Ks < 0.55 for intersyntenic blocks of loquat–*G. trifoliata*, apple–*G. trifoliata* and pear–*G. trifoliata*). Gene pairs in the syntenic blocks were then processed using the "OMG!" program (79) with default parameters. A total of 18,764 nonredundant identified homologous gene sets were produced according to the following conditions: minimum zero genes and maximum two genes in apple, loquat, and pear; minimum zero genes and maximum one gene in *G. trifoliata*; and each gene family contains at least two genes. Each of these sets represented one "candidate gene" in the reconstructed ancestral chromosomes. Based on the nonredundant homologous gene sets from OMG, a program was written using the maximum weight algorithm to construct the Maleae ancestral genome. First, we identified all the gene adjacencies in loquat, apple, pear, and *G. trifoliata* by only considering genes in the nonredundant homologous gene sets. Each adjacency was weighted according to the numbers of homologous copies, producing an optimal set of 1,284 contigs, containing 9,388 genes, with an average of 7.3 genes per contig. Second, using the adjacencies of contigs in each genome as an input, the program was rerun to obtain an optimal of 169 scaffolds containing 9,055 genes. Each scaffold has 53.6 genes. Finally, using scaffold's adjacencies in each genome as input, the program was rerun to obtain 30 superscaffolds, which contained 9,237 genes, with an average of 307.9 genes per superscaffold. To minimize the number of crossing lines, the superscaffold order in the chromosomes was rearranged to obtain nine ancestral chromosomes, which contained 9,115 genes, with each chromosome containing 1,012.7 genes.

Genome Mining and Phylogenetic Analysis. HMMER 3 (80) was used to identify OSC homolog genes. PF13243 (targeting N terminal of OSC) and PF13249 (targeting C terminal of OSC) were used to query for OSCs using the cut_tc (trusted cutoff) option. CYP716 sequences were screened by BLASTP using *Medicago truncatula* CYP716A12 (GenBank accession: DQ335781) (81) as query, with an E-value cutoff at 1e−5, protein identity ≥40 and bit score ≥100. Accession numbers for characterized OSCs and cytochrome P450s are listed in *SI Appendix*, Table S6. Phylogenetic analyses were performed using an ML method with the GTRGAMMA substitution model implemented in RaxML (76), and bootstrap values at the branch nodes were calculated from 1,000 replications.

Gene Correlation Analysis. WGCNA was used to identify associated gene groups. For transcriptome analysis, 10 and 9 tissues for loquat and *G. trifoliata* were used, respectively (*Data Availability*). We inferred a weighted undirected coexpression network for different tissues of loquat and *G. trifoliata* separately using the WGCNA (82) package in R with a soft thresholding power of 16 and 18. Groups of closely correlated genes (form a module) were identified by clustering genes based on the topological overlap matrix and resulted dendrogram produced by cutreeDynamic method (parameters: deepSplit = 2, pamRespectsDendro = FALSE, minModuleSize = 50). Nonmodule genes were grouped as an artificial module (gray). Initial groups with similar gene expression profiles (eigengene correlation ≥0.75) were merged. We used exportNetworkToCytoscape in the WGCNA package to create an edge file with weight value between two genes. The weight value referred to the connection strength between any two genes, with the higher value referring to a stronger expression correlation.

Isolation of OSCs and P450s for Triterpenes Biosynthesis. The cDNAs for targeted genes were first cloned into pDNOR-207 entry vector by a BP reaction (Gateway; Invitrogen). Constructs were verified by Sanger sequencing. pEAQ-HT-DEST-1 was used as a destination vector for an attL-attR sites recombination (LR) reaction. The primers used in this study were listed in *SI Appendix*, Table S7. Transient expression of OSCs and P450s in *N. benthamiana* were performed as previously reported (45). Expression vectors were introduced into *Agrobacterium tumefaciens* strain LBA4404. The strains harboring expression constructs were freshly grown on Lysogeny Broth medium with antibiotic selection (50 $\mu g \cdot mL^{-1}$ kanamycin, 50$\mu g \cdot mL^{-1}$ rifamycin & 50 $\mu g \cdot mL^{-1}$ streptomycin) and incubated at 28 °C with 220 rpm until an OD_{600} = 2 for about 16 h. LBA4404 cells were pelleted by centrifuging at 4,500 *g* for 20 min, and supernatants were discarded. The pellets were then resuspended in freshly made MMA buffer (10 mM MgCl2, 10 mM MES/KOH pH5.6, 150 μM acetosyringone) and diluted to OD_{600} = 0.2. For combinatorial assay, strains harboring different constructs were mixed and infiltrated by a syringe without a needle. Plants were grown to about six leaves stage. Leaves were collected 6 d after infiltration.

Triterpene Analysis. Leaves were dried in a Frerzone 12L lyophilizer (Labconco) before being ground into fine powder. For triterpene analysis, 5 mg powder was taken and saponified by 500 μL saponification solution ethanol/H_2O/KOH pellets in 9:1:1 (volume[vol]/vol/weight) at 70 °C for 1 h. The ethanol was removed by evaporation for 1 h at 70 °C. The triterpenes were extracted with 1 mL ethyl acetate/H_2O in 1:1 (vol/vol). After centrifugation at 16,000 *g* for 1 min, the supernatants were collected and dried under N_2. The extracts were resuspended in 20 μL Methoxyamine hydrochloride-Pyridine mixture (20 $mg \cdot ml^{-1}$; Sigma-Aldrich) and incubated at 37 °C for 2 h. A total of 30 μL *N*-Methyl-*N*-(trimethylsilyl) trifluoroacetamide was added into the mixture and incubated at 37 °C for another 30 min before GC-MS analysis. The triterpenes were analyzed on an Agilent 7890A/5975C GC system equipped with an HP-5M column (30 m × 0.25 mm × 0.25 μm). Helium at a flow rate of 1 $mL \cdot min^{-1}$ was used as carrier gas. We injected 1 μL of each extraction with a GC inlet at 250 °C. The oven temperature program began from 170 °C (held for 2 min) to 290 °C (held for 4 min) at a speed of 6 °C min^{-1} and switched to 320 °C (held 15 min) at a rate of 20 °C min^{-1}. For metabolite identification, full mass spectra were generated by scanning within the *m/z* range of 60 to 800. Triterpenes were monitored by comparing both the retention time and mass spectra with the authentic standards of BA (CAS no. 559-70-6), AA (CAS no. 508-04-3), OA (CAS no. 508-02-1), UA (CAS no. 77-52-1), MA (CAS no. 4373-41-5), and CA (CAS no. 4547-24-4). The internal standard (coprostanol, CAS no. 360-68-9) was purchased from Sigma-Aldrich.

Quantification of Triterpenes. For triterpene quantifications of the Rosaceae species, authentic standards at concentrations range from 5 $\mu g \cdot ml^{-1}$ to 100 $\mu g \cdot ml^{-1}$ (5, 10, 20, 30, 40, 50, 60, 70, and 100 $\mu g \cdot ml^{-1}$) were used to obtain calibration curves based on the peak area of each standard. Calibration curves and coefficients of each standard are as follows: BA: $y = 5.1338x + 4.7671$, $R^2 = 0.9977$; AA: $y = 2.382x + 38.371$, $R^2 = 0.9972$; OA: $y = 3.1334x + 7.8062$, $R^2 = 0.9952$; UA: $y = 3.5107x + 8.8741$, $R^2 = 0.9969$; MA: $y = 5.7214x + 8.6611$, $R^2 = 0.9988$; and CA: $y = 5.6682x + 7.7736$, $R^2 = 0.9968$. For relative quantification of triterpenes in tobacco leaves, coprostanol at 5 $\mu g \cdot ml^{-1}$ was used as an internal standard. The ratios of products were calculated as percentages of a peak area relative to the total identified peak areas.

Data Availability. The genome (loquat and *G. trifoliata*) and transcriptome data have been deposited in the China National GeneBank DataBase (CNGB) Nucleotide Sequence Archive (https://db.cngb.org/cnsa) with accession number CNP0001531. The cytochrome P450s that were functionally characterized in this study were formally assigned as CYP716A125_Eriobotrya_japonica, CYP716AV6_Eriobotrya_japonica, CYP716C17_Eriobotrya_japonica, CYP716C15_Eriobotrya_japonica, and CYP716A125_Gillenia trifoliata according to procedures of the Cytochrome P450 homepage (83). All other study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. This work was supported by the Ministry of Agriculture and Rural Affairs of People's Republic of China Industry Technology Special Project 201003073 (Shunquan Lin), the Universities Service Haixi Construction Key Program of Fujian Province 2008HX02 (J.W.), the National Natural Science Foundation of China 31901973 (W.S.), the National Key Research and Development Program 2019YFD1000200 (X.Y.), and the Natural Science Foundation of Fujian Province 2017J01644 (S.L.). Zhenhua Liu's program is

sponsored by Shanghai Pujiang Program. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant no. ERC-StG 679056 HOTSPOT), via a grant to L.Y. We thank Prof. Jun Wu (Nanjing Agricultural University) for providing the transcriptomic data of pear. We thank Profs. Yaoguang Liu and Qiang Xu for their critical reading and suggestions. We thank Prof. David Nelson (The University of Tennessee) for assigning names to the characterized P450s in this study.

1. J. K. Weng, The evolutionary paths towards complexity: A metabolic perspective. *New Phytol.* **201**, 1141–1149 (2014).
2. Z. Liu *et al*., Evolutionary interplay between sister cytochrome P450 genes shapes plasticity in plant metabolism. *Nat. Commun.* **7**, 13026 (2016).
3. N. Panchy, M. Lehti-Shiu, S.-H. Shiu, Evolution of gene duplication in plants. *Plant Physiol.* **171**, 2294–2316 (2016).
4. X. Qiao *et al*., Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* **20**, 38 (2019).
5. J. W. Clark, P. C. J. Donoghue, Whole-genome duplication and plant macroevolution. *Trends Plant Sci.* **23**, 933–945 (2018).
6. Y. Van de Peer, S. Maere, A. Meyer, The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732 (2009).
7. S. Liu *et al*., The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nat. Commun.* **5**, 3930 (2014).
8. T. Unver *et al*., Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E9413–E9422 (2017).
9. P. P. Edger *et al*., The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8362–8366 (2015).
10. D. J. Kliebenstein, A role for gene duplication and natural variation of gene expression in the evolution of metabolism. *PloS One* **3**, e1838 (2008).
11. B. M. Moore *et al*., Robust predictions of specialized metabolism genes through machine learning. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2344–2353 (2019).
12. S. Lin, R. H. Sharpe, J. Janick, Loquat: Botany and horticulture. *Hortic. Rev.* **23**, 233–276 (1999).
13. Y. Liu, W. Zhang, C. Xu, X. Li, Biological activities of extracts from loquat (Eriobotrya japonica Lindl.): A review. *Int. J. Mol. Sci.* **17**, 1983 (2016).
14. S. Z. Li, *Compendium of Materia Medica* (People's Med. Publ. House Beijing, China, 1578) (In Chinese).
15. Y. Zhang *et al*., A dual effect of ursolic acid to the treatment of multiple sclerosis through both immunomodulation and direct remyelination. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9082–9093 (2020).
16. X. Li *et al*., Cell-penetrating corosolic acid liposome as a functional carrier for delivering chemotherapeutic drugs. *Acta Biomater.* **106**, 301–313 (2020).
17. H. Tan *et al*., Ursolic acid isolated from the leaves of loquat (Eriobotrya japonica) inhibited osteoclast differentiation through targeting exportin 5. *J. Agric. Food Chem.* **67**, 3333–3340 (2019).
18. Y. Xiang *et al*., Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* **34**, 262–281 (2017).
19. K. R. Robertson, J. B. Phipps, J. R. Rohrer, E. Claire, P. G. Smith, A synopsis of genera in Maloideae (Rosaceae). *Bot. Source Syst.* **16**, 376–394 (1991).
20. N. Daccord *et al*., High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
21. X. Sun *et al*., Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* **52**, 1423–1432 (2020).
22. J. Wu *et al*., The genome of the pear (Pyrus bretschneideri Rehd.). *Genome Res.* **23**, 396–408 (2013).
23. V. Shulaev *et al*., The genome of woodland strawberry (Fragaria vesca). *Nat. Genet.* **43**, 109–116 (2011).
24. I. Verde *et al*.; International Peach Genome Initiative, The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**, 487–494 (2013).
25. B. J. Walker *et al*., Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **9**, e112963 (2014).
26. R. Velasco *et al*., The genome of the domesticated apple (Malus × domestica Borkh.). *Nat. Genet.* **42**, 833–839 (2010).
27. O. Raymond *et al*., The Rosa genome provides new insights into the domestication of modern roses. *Nat. Genet.* **50**, 772–777 (2018).
28. Q. Zhang *et al*., The genome of Prunus mume. *Nat. Commun.* **3**, 1318 (2012).
29. R. C. Evans, C. S. Campbell, The origin of the apple subfamily (Maloideae; Rosaceae) is clarified by DNA sequence data from duplicated GBSSI genes. *Am. J. Bot.* **89**, 1478–1484 (2002).
30. N. De Tommasi *et al*., Constituents of Eriobotrya japonica. A study of their antiviral properties. *J. Nat. Prod.* **55**, 1067–1073 (1992).
31. National Commission of Chinese Pharmacopoeia, *China Med* (Sci. Technol. Press, Beijing, 2015), pp. 204.
32. D. W. Christianson, Structural and chemical biology of terpenoid cyclases. *Chem. Rev.* **117**, 11570–11648 (2017).
33. I. Verde *et al*., The peach v2.0 release: High-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genom.* **18**, 225 (2017).
34. C. M. Andre *et al*., Multifunctional oxidosqualene cyclases and cytochrome P450 involved in the biosynthesis of apple fruit triterpenic acids. *New Phytol.* **211**, 1279–1294 (2016).
35. K. Miettinen *et al*., The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis. *Nat. Commun.* **8**, 14153 (2017).
36. R. C. Sandeep, R. C. Misra, C. S. Chanotiya, P. Mukhopadhyay, S. Ghosh, Oxidosqualene cyclase and CYP716 enzymes contribute to triterpene structural diversity in the medicinal tree banaba. *New Phytol.* **222**, 408–424 (2019).
37. S. Yasumoto, H. Seki, Y. Shimizu, E. O. Fukushima, T. Muranaka, Functional characterization of CYP716 family P450 enzymes in triterpenoid biosynthesis in tomato. *Front. Plant Sci.* **8**, 21 (2017).
38. F. M. Afendi *et al*., KNApSAcK family databases: Integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.* **53**, e1 (2012).
39. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, New York, 1970).
40. L. E. Flagel, J. F. Wendel, Gene duplication and evolutionary novelty in plants. *New Phytol.* **183**, 557–564 (2009).
41. J.-K. Weng, R. N. Philippe, J. P. Noel, The rise of chemodiversity in plants. *Science* **336**, 1667–1670 (2012).
42. L. Mao *et al*., Genomic evidence for convergent evolution of gene clusters for momilactone biosynthesis in land plants. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 12472–12480 (2020).
43. Q. Wang, M. L. Hillwig, Y. Wu, R. J. Peters, CYP701A8: A rice ent-kaurene oxidase paralog diverted to more specialized diterpenoid metabolism. *Plant Physiol.* **158**, 1418–1425 (2012).
44. Z. Liu *et al*., Formation and diversification of a paradigm biosynthetic gene cluster in plants. *Nat. Commun.* **11**, 5354 (2020).
45. Z. Liu *et al*., Drivers of metabolic diversification: How dynamic genomic neighbourhoods generate new biosynthetic pathways in the brassicaceae. *New Phytol.* **227**, 1109–1123 (2020).
46. J.-K. Weng, Y. Li, H. Mo, C. Chapple, Assembly of an evolutionarily new pathway for α-pyrone biosynthesis in Arabidopsis. *Science* **337**, 960–964 (2012).
47. S. Chen *et al*., CYP79F1 and CYP79F2 have distinct functions in the biosynthesis of aliphatic glucosinolates in Arabidopsis. *Plant J.* **33**, 923–937 (2003).
48. G. Carrier *et al*., An efficient and rapid protocol for plant nuclear DNA preparation suitable for next generation sequencing methods. *Am. J. Bot.* **98**, e13–e15 (2011).
49. W. Zhu *et al*., Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid. *Genome Biol.* **18**, 157 (2017).
50. J. Doležel, J. Greilhuber, J. Suda, Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* **2**, 2233–2244 (2007).
51. S. Koren *et al*., Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
52. H. Liu, S. Wu, A. Li, J. Ruan, SMARTdenovo: A de novo assembler using long noisy reads. *Gigabyte*, 10.46471/gigabyte.15 (2021).
53. Y. Chen *et al*., SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* **7**, 1–6 (2018).
54. N. C. Durand *et al*., Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
55. S. S. P. Rao *et al*., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
56. B. J. Haas *et al*., De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
57. M. Tarailo-Graovac, N. Chen, Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **4**, 10 (2009).
58. W. Bao, K. K. Kojima, O. Kohany, Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
59. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
60. T. M. Lowe, S. R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
61. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
62. E. P. Nawrocki *et al*., Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137 (2015).
63. C. Holt, M. Yandell, MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf.* **12**, 491 (2011).
64. A. D. Johnson *et al*., SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
65. M. Stanke *et al*., AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435-9 (2006).
66. E. M. Zdobnov, R. Apweiler, InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
67. A. Conesa *et al*., Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
68. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
69. W. J. Kent, BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
70. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

71. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinf.* **12**, 323 (2011).
72. C. Chen *et al*., TBtools-an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* **13**, 1194–1202 (2020).
73. L. Li, C. J. Stoeckert Jr, D. S. Roos, M. C. L. Ortho, OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
74. R. C. Edgar, MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinf.* **5**, 113 (2004).
75. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
76. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
77. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
78. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
79. C. Zheng, K. Swenson, E. Lyons, D. Sankoff, "OMG! Orthologs in multiple genomes - competing graph-theoretical formulations" in *Proceedings of the Eleventh International Workshop on Algorithms in Bioinformatics (WABI)*, T. M. Przytycka, M.-F. Sagot, Eds. (Springer, 2011), pp. 364–375.
80. S. C. Potter *et al*., HMMER web server: 2018 update. *Nucleic Acids Res.* **46** (W1), W200–W204 (2018).
81. M. Carelli *et al*., Medicago truncatula CYP716A12 is a multifunctional oxidase involved in the biosynthesis of hemolytic saponins. *Plant Cell* **23**, 3070–3081 (2011).
82. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinf.* **9**, 559 (2008).
83. D. R. Nelson, The cytochrome p450 homepage. *Hum. Genom.* **4**, 59–65 (2009).